PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : H04L 29/06, 12/56 — A1

(11) International Publication Number: WO 99/23799

(43) International Publication Date: 14 May 1999 (14.05.99)

(21) International Application Number: PCT/GB98/03274

(22) International Filing Date: 3 November 1998 (03.11.98)

(30) Priority Data:
97308790.1 3 November 1997 (03.11.97) EP

(71) Applicant *(for all designated States except US)*: BRITISH TELECOMMUNICATIONS PUBLIC LIMITED COMPANY [GB/GB]; 81 Newgate Street, London EC1A 7AJ (GB).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: HODGKINSON, Terence, Geoffrey [GB/GB]; 46 Melton Grange Road, Melton, Woodbridge, Suffolk IP12 1SD (GB). CARTER, Simon, Francis [GB/GB]; 5 Moorfield Road, Woodbridge, Suffolk IP12 4JN (GB). O'NEILL, Alan, William [GB/GB]; 2 Rachaels Court, 36 Cemetery Road, Ipswich, Suffolk IP4 2JA (GB). WHITE, Paul, Patrick [GB/GB]; 82 Ringlow Park Road, Swinton, Manchester M27 0HB (GB).

(74) Agent: NASH, Roger, William; BT Group Legal Services, Intellectual Property Dept., Holborn Centre, 8th floor, 120 Holborn, London EC1N 2TE (GB).

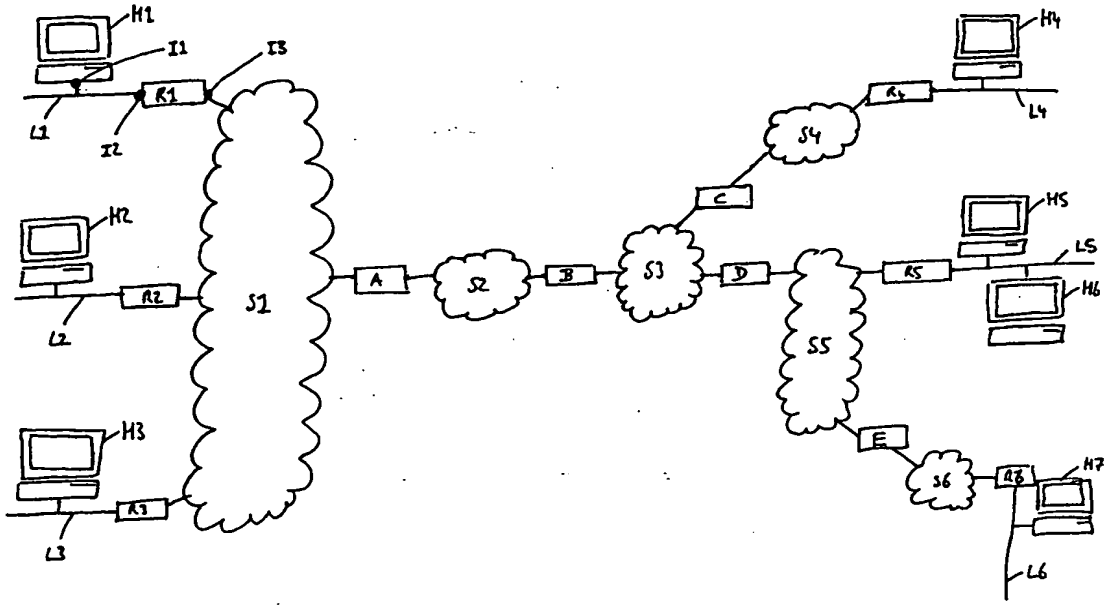
(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published

With international search report.

Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: PACKET NETWORK



(57) Abstract

A method of reserving resources in an internet is disclosed. The method provides an improved process for use in relation to large scale shared–tree multicast environments since its use results in a reduction in path state in the routers (A–E, R1–R6) in the internet. The method involves the sending of path characteristics upstream from receivers (H1–H7) to senders (H1–H7), the routers in between combining path characteristics from different sources downstream of them. Reservations are subsequently made on the basis of the combined path characteristic data, the nature of the sender's traffic and the end–to–end quality of service required by the sender.

PACKET NETWORK

The present invention relates to a method of reserving resources in a packet network. It has particular utility in relation to providing an internet offering Quality of Service guarantees.

Methods of reserving resources in an internet are well known. The method which is currently best supported in the Internet is that defined by the Resource Reservation Protocol (RSVP).

There is a desire to enable an internet to provide real-time communication as is, for example, required if telephone conversations or video-conferences are to be conducted over it. In this regard, two qualities of service that might be provided have been specified by the following documents (which are incorporated herein by reference):

(1) S. Schenker, C.Partridge, R.Guerin. Specification of Guaranteed Quality of Service, Request For Comments, September 1997, RFC 2212; and

(2) J. Wroclawski. Specification of the Controlled-Load Network Element Service, Request For Comments, September 1997, RFC 2211.

Essentially, Guaranteed Service as defined in the first document allows a user to specify an upper bound on the time taken for his message to reach a recipient, whereas Controlled Load Service offers a service qualitatively similar to that provided by the internet when it is only lightly loaded. Operating an internet in accordance with the RSVP protocol allows the provision of real-time communication. The provision of such communication to the above-mentioned Quality of Service classes when operating an internet in accordance with the RSVP protocol is discussed in the following document (also incorporated herein by reference):

operating said one or more intermediate nodes to:

combine one or more parameters of received reverse-routed packets from different receivers to generate combined reservation influencing parameters;

store said combined parameters; and

send a reverse-routed packet containing said combined parameters further along said route towards said sender hosts;

operating said sender hosts to send a resource reservation packet to one or more receivers back along said route; and

operating said one or more intermediate nodes, responsive to said reservation packet, to reserve resources in accordance with reservation influencing parameters stored at that node.

Because intermediate nodes examine reservation influencing data sent by receivers, and reserve resources accordingly, the present invention better tailors a resource reservation to receivers' requirements. Furthermore, the present invention achieves this without sacrificing the faster resource reservation associated with sender initiated resource reservation protocols.

Also, by combining path characteristic data in this way, the amount of path characteristic data sent between nodes in a network where a sender sends traffic to many receivers (i.e. where the sender is multicasting) is reduced.

In some embodiments, said one or more reservation influencing parameters include an indication of a desired quality of service class, and said one or more intermediate nodes are operable to:

update said parameters to represent the highest quality of service class requested by downstream receivers; and

reserve resources in accordance with the resource reservation process associated with said highest quality of service class.

In this way, it is possible to support different quality of service classes within a single one-to-many or many-to-many communication. For example, some receivers might request a quality of service in accordance with the Guaranteed Service mentioned above, whereas others might only require a quality of service in

Figure 2 is a schematic illustration of the contents of a reservation setting packet used in a first embodiment of the present invention;

Figure 3 is a schematic illustration of the contents of an upstream-routed worst path characteristics per node packet used in the first embodiment;

Figure 4 is a schematic illustration of worst path characteristics per path data stored in a node which operates in accordance with the first embodiment; and

Figure 5 is a schematic illustration of worst path characteristics per interface data stored in a node which operates in accordance with the first embodiment.

Figure 1 shows seven computers (H1 to H7), each of which is connected to a Local Area Network (L1 to L6) which also includes a gateway router (R1 to R6). Two of the computers H5, H6 are connected to a single Local Area Network (LAN), the other five computers are connected to respective LANs. The gateway routers (R1 to R7) are interconnected by means of a network which comprises a number of subnets (S1 to S6) connected to one or more other subnets by intermediate routers (A to E).

A first subnet S1 connects three of the gateway routers (R1to R3) to a first intermediate router A, which is in turn connected via a second subnet S2 to a second intermediate router B. A third subnet S3 connects the second intermediate router to third and fourth intermediate routers C,D. The third intermediate router C is connected via a fourth subnet S4 to the fourth gateway router R4. The fifth subnet S5 connects the fourth intermediate router D to both the fifth intermediate router E and the fifth gateway router R5. The fifth intermediate router E is connected to the sixth gateway router R6 via a sixth subnet S6.

The computers (H1 to H7), routers (R1 to R7, A to E) and subnets (S1 to S6) form a portion of an internet. The internet includes a number of other computers (not shown) connected to the LANs (L1 to L6) and might include further subnets, routers and LANs. The internet is operable to allow the computers (H1 to H7) to

for a message from, say, first host H1 to the other hosts, the interface I1 between the first router R1 and the first LAN L1 is known as an upstream interface, and the interface I2 between the first router R1 and the first subnet S1 is known as a downstream interface.

Broadly, according to a first embodiment of the present invention, the portion of the internet serving the seven hosts (H1 to H7) is operable in accordance with the TCP/IP network architecture but is additionally operable in accordance with the following procedure to allow each of the hosts to reserve resources of the internet for communication with the other hosts. By having the hosts reserve appropriate resources the characteristics of the communication paths to the other hosts can be controlled. For example, the delay in the communication can be minimised or reduced below a predetermined bound.

In the embodiment, each host occasionally sends a Reservation Packet (hereinafter an RES packet) to all the other hosts. Part of the purpose of this packet is to provide each node with the address of the downstream interface of the node directly upstream on the source-based tree which the message follows to each of the other hosts. This address is known as a previous hop address. Those skilled in the art will recognise that a similar function is performed by the PATH messages of the RSVP protocol.

Once a recipient host has received at least one RES packet, it will start sending Backwards Control packets (hereinafter BWDC packets) towards the sending host. Normally, the recipient host will send one such packet in a predetermined refresh period (though in certain circumstances (to be described below), it will send more than one).

Each of the nodes reads any BWDC packet it receives and stores worst path characteristic values included in the packet in its memory. The stored values are then used to update data representing the worst per interface path characteristics (stored data like this is known as a 'state' entry to those skilled in the art). A similar worst per interface state entry is made in relation to each of any other downstream interfaces of the node. The state entries are then compared and the

- **session** - this is identical to the session field defined by the RSVP protocol - it includes the destination address (a multicast address for the multicast group) for the flow (i.e. one or more messages) to which the RES packet relates and other parameters relating to the flow. The nature of these parameters is known to those skilled in the art

- **Sender Template** - this is identical to the corresponding field defined by the RSVP protocol - i.e. it is a filter specification identifying the sending host. It contains the IP address of the sending host and optionally the sending host port being used.

- **Traffic specification (Tspec)** this is identical to the corresponding field defined by the RSVP protocol describing the sending host's traffic characteristics using the following token bucket representation.

  p = peak rate of flow (bytes/second)

  b = bucket depth (bytes)

  r = token bucket rate (bytes/second)

  m = minimum policed unit (bytes)

  M = maximum datagram size (bytes)

- **previous hop (Phop)** - this is identical to the corresponding field defined by the RSVP protocol i.e. it is an object including the previous hop address.

- **timestamp** field - this is stamped with the time of the local node clock just before being forwarded to the next node(s) down the distribution tree.

- **end-to-end delay** field - this gives the current delay from when a packet was transmitted by the sending host until it is due to arrive at the upstream interface of the next node.

- **CRTs** field(2 bits) - this identifies the ceiling reservation type of the sending host application. 11 indicates guaranteed service, 10 indicates controlled-load, and 00 indicates best-effort. 01 is currently unspecified although may at some time

The other information contained in the BWDC packet includes:

- **session** - this is identical to the session field defined above for the corresponding RES packet.

- **downstream hop** object - this is identical to RSVP next hop object - i.e. it gives the address of the upstream interface of the node directly downstream that sent the packet.

- **timestamp** field which is stamped with the time of the local node clock just before being sent to the node directly upstream.

- **timedeltaprev** field - this is filled in with the stored value of timedeltaprev (whose value is explained below) just before being sent to the node directly upstream.

- **CRTr** field(2 bits). This field indicates the recipient host application ceiling reservation type. The mapping between the values of this field and the reservation types they represent are the same as for the CRTs field in the RES message.

- **Worst Case Delay** field. This equals the maximum data packet propagation delay measured between the upstream interface of the node from which the BWDC packet was sent and each recipient host downstream of that node.

- **path MTU** field. This equals the minimum pathMTU value between the upstream interface of the node from which the BWDC packet was sent and each recipient host downstream of that node.

- **Worst Case Ctot** field - this is as defined in the Guaranteed Service specification - i.e. it equals the maximum accumulated Ctot value along the paths between the upstream interface of the node from which the BWDC packet was sent and each recipient host downstream of that node.

then set in accordance with the node's local clock and the RES packet is sent onwards to the next node(s) along the source-based tree towards the recipient hosts. This process is repeated until each of the nodes involved store a parameter (timedeltaprev) representing the propagation delay over the hop directly upstream of them (when traversed in the downstream direction).

Once the RES packets have propagated through the nodes of the source-based tree, each of the recipient hosts sends a BWDC packet towards the sending hosts. The initial values placed in the packet by the hosts are determined as follows. The path MTU, and path bandwidth correspond to the characteristics of the LAN to which the sending host is attached. The host inserts its IP address in the downstream node field and sets CRTr in accordance with the quality of service it requires. The worst case path characteristics are all set to zero, as is the bottleneck flag. The excess delay field is unused at this stage. The timedeltaprev parameter is assigned to the timedeltaprev field.

On receipt of a BWDC packet a node carries out timing operations and worst per node path characteristic determining operations as described below.

The timing operations involve the node first reading the timedeltaprev field in the packet. It will be realised that this records the propagation delay (experienced by the RES packet) over the hop downstream from the node (when traversed in the downstream direction). By also reading the timestamp field of the BWDC packet and the local clock a value for the propagation delay over the same hop in the other direction is obtained. It is assumed that the delay over the downstream hop is the same in either direction. Hence, by taking the average of these two delays a value of the propagation delay independent of any discrepancy between the node clocks is obtained. This average delay for the downstream hop is stored by the node as a parameter 'dnext'.

The path characteristic monitoring operations are as follows.

Upon arrival of an BWDC packet, the node first checks for the existence of a worst per path state entry relating to the current session and to the downstream path

As shown in Figure 5, the worst per interface state entry contains the same fields as the worst per path state entry save that it lacks the downstream node field. The values assigned to the session, sender address field and downstream interface fields are those of the worst per path state entries. The values for the other fields are calculated as follows.

- Worst Case Ctot = MAX{Worst Case $Ctot_i$}
- Worst Case Dtot = MAX{Worst Case $Dtot_i$}
- Worst Case Delay = MAX{Worst Case $Delay_i$}
- path bandwidth = MAX{path $bandwidth_i$}
- pathMTU = MIN{$pathMTU_i$}
- CRTr = MAX{CRT_i}
- dnext = MAX{$dnext_i$}

where the subscript i takes values from 1 to the number of nodes sharing the downstream interface. It will be seen that the values Worst Case Ctot, Worst Case Dtot and Worst Case Delay that are stored in the worst per interface state entry are worst case path characteristic values for the downstream interface. It is possible that the values relate to different paths from the interface. The bottleneck flag is set to one if any of the per path state entries have it set to one.

Having created a worst per interface state entry for each of its downstream interfaces, the node then generates a BWDC packet containing worst per node data at least once every refresh period. The parameters to be included in the worst per node data are calculated as follows

- Worst Case Ctot = MAX{Worst Case $Ctot_n + Clocal_n$} where $Clocal_n$ is the value of Ctot between the downstream interface on which the BWDC packet arrived and the upstream interface (determined by the session) of the node;

At the first node downstream along the source-based tree, the queuing delay to be imposed on the flow in relation to each of the one or more downstream hops involved in the current session. This is determined in accordance with the equation below:

Equation (1)

Qdelay = desired-bound - accumulated-bound - Worst Case Delay

The Worst Case Delay value equals the sum of the corresponding field and the dnext parameter of the worst per interface state entry at the current node. At the first node the accumulated bound will normally be zero or negligible. The Qdelay value represents an estimate of the total queuing delay that can be tolerated over the remainder of the path to a recipient host.

Processing similar to that carried out by receivers operating in accordance with the RSVP protocol is then used to calculate a bandwidth to be reserved on each of the downstream hops in order to stay 'on-course' for the desired bound. This calculation will often be an overestimate because it may be that the individual worst per interface parameters relate to different paths from that interface. Those skilled in the art will realise that the estimate is calculated using the following equations:

Equation 2

$$Qdelay_{end2end} = \frac{(b-M)(p-R)}{R(p-r)} + \frac{(M+Ctot)}{R} + Dtot \qquad \text{(case p > R >= r)}$$

Equation 3

$$Qdelay_{end2end} = \frac{(M+Ctot)}{R} + Dtot \qquad \text{(case R >= p >= r)}$$

The parameters M, p, b and r are obtained from the corresponding fields of the RES message (the meaning of those symbols is as set out above in relation to the RES message). The values of Ctot and Dtot are a sum of the values from the

unless re-shaping of the sending hosts traffic is carried out at the downstream interface, in which case both CSum and Dsum are set to zero.

Also, the Qosvoid field is set to one if either

a) a Guaranteed Service reservation attempt fails to reserve a bandwidth at least as great as the token bucket rate of the traffic specification; or

b) a Controlled Load reservation attempt fails.

If a node receives a RES packet with Qosvoid set to one, then if MIN{CRTs, CRTr} is 10 or 11 it attempts to secure a Controlled Load reservation.

Once updating of the fields is complete the timestamp field is (as described above) set equal to the local clock before forwarding the RES packet to each next hop down the routing tree.

Similar processing is carried out at each of the subsequent nodes, the increase of the accumulated bound by the installed queuing delay resulting in the value of Qdelay obtained from equation (1) continuing to be an estimate of the total queuing delay that can be tolerated for the remainder of the path to a recipient host. Assuming the messages from the host to be in accordance with the declared traffic specification, and provided each of the nodes is able to reserve the calculated bandwidth R, the above-described processing at the nodes will be effective to enable messages from the sending hosts to be delivered within the desired delay bound to all of the recipient hosts.

If any node is unable to reserve the bandwidth as calculated above then (subject to policy configurations at the node) as much bandwidth as possible is reserved. If, however, the amount of bandwidth reserved is less than the value of the token bucket rate field of the RES packet then Guaranteed Service is not offered and this is indicated by setting the delayvoid, lossvoid and Qosvoid flags to 1.

If the bandwidth reserved is more than the value of the token bucket rate field of the RES packet, then the accumulated bound is compared to the desired bound.

In other embodiments the resource reservation might be calculated on the basis of combined data at the node where the combination takes place.

A second embodiment is similar to the first embodiment but uses a shared tree protocol. In the second embodiment the sender address field (in the BWDC packet, the worst per path and worst per interface state entries) is not required. This is because the interfaces out of which the RES packet are to be sent are determinable by the node on the basis of the session parameter and the incoming interface. Since, when using a shared tree the worst per interface state entries will relate to the multicast group rather than a specific sender to the multicast group, the number of worst per interface state entries in each node is only as great as the number of outgoing interfaces from the node.

It will be seen how, in multi-sender environments, the second embodiment reduces the amount of worst per interface state entries required in comparison with known reservation protocols.

In the above embodiments, the worst-case merging of C terms (e.g. Ctot), D terms (e.g. Dtot) and link propagation delay was carried out for each term independently. This results in an overly conservative local bandwidth reservation. In preferred embodiments, a rate independent delay parameter which includes both the D term and the link propagation delay is used. This might be done by taking the forwarded value of D as the value from the worst case rate independent delay parameter rather than simply the maximum D value from each path.

It will be seen how the above embodiments enable a router to find the highest quality of service requested by any one of the downstream receivers. Similar considerations apply to path characteristic data and other reservation influencing parameters. It is the combination of such parameters at intermediate nodes in the network that allows a more flexible resource reservation to be provided for multicast communication in a packet network.

5. A method according to any preceding claim wherein said hosts communicate in accordance with a shared tree routing algorithm.

6. A packet network node comprising:

means for combining one or more parameters of received reverse-routed packets to generate combined reservation influencing parameters;

means for storing said combined parameters;

means for sending a reverse routed packet containing said combined parameters further along said route to said sender hosts; and

means for reserving resources in accordance with combined reservation influencing parameters stored at that node responsive to a reservation packet.

7. A method of reserving resources in a packet network comprising a plurality of hosts and one or more interconnecting nodes, said method comprising the steps of:

operating one or more receiver hosts to send path characteristic data packets, containing one or more path parameters, along a route via one or more of said interconnecting nodes to one or more sender hosts;

operating said one or more intermediate nodes to:

process one or more parameters of received path characteristic data packets to generate updated path characteristic parameters;

store said updated parameters; and

send a path characteristic data packet containing said updated parameters further along said route to said sender hosts

operating said sender hosts to send a resource reservation packet to one or more receivers back along said route; and

operating said one or more intermediate nodes, responsive to said reservation packet, to reserve resources in accordance with path characteristic data stored at that node.

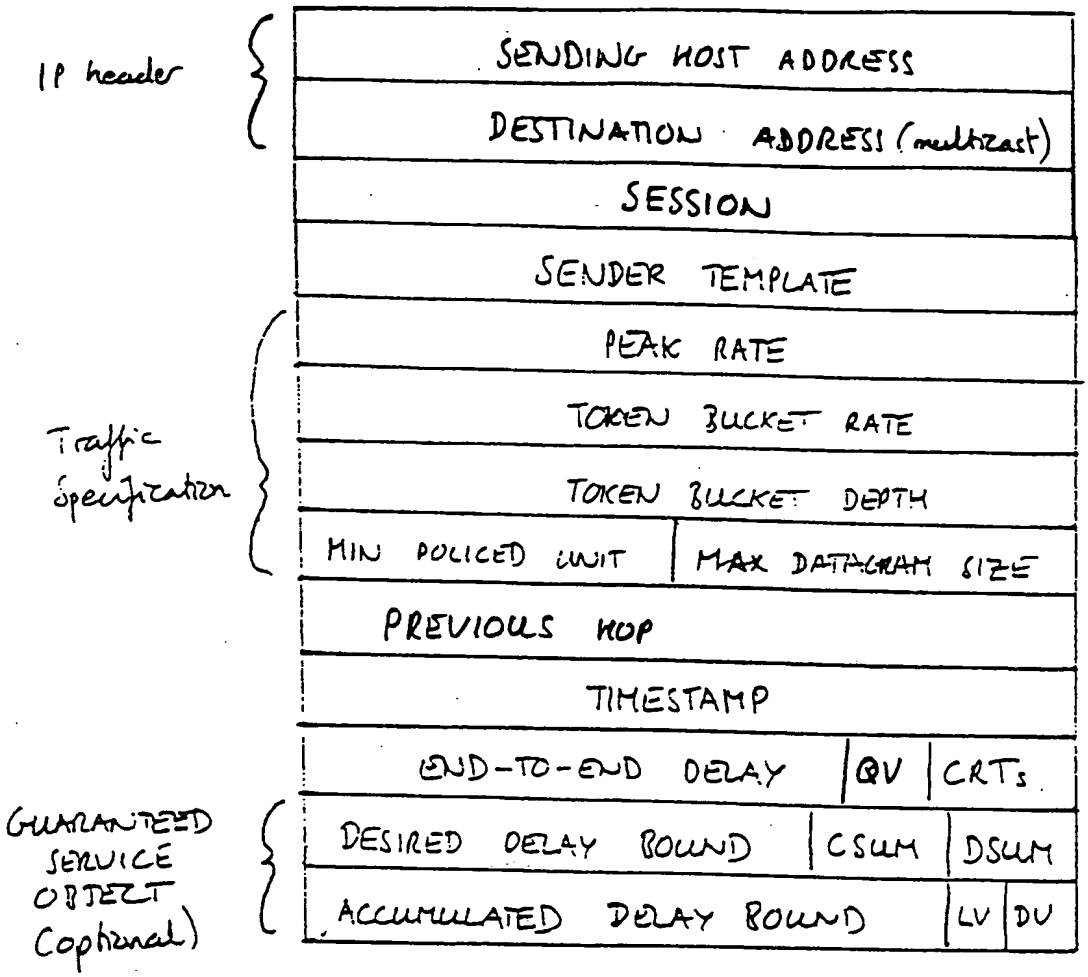

IP header
SENDING HOST ADDRESS
DESTINATION ADDRESS (multicast)
SESSION
SENDER TEMPLATE
Traffic Specification
PEAK RATE
TOKEN BUCKET RATE
TOKEN BUCKET DEPTH
MIN POLICED UNIT
MAX DATAGRAM SIZE
PREVIOUS HOP
TIMESTAMP
END-TO-END DELAY
QV
CRTs
GUARANTEED SERVICE OBJECT (Optional)
DESIRED DELAY BOUND
CSUM
DSUM
ACCUMULATED DELAY BOUND
LV
DV

Figure 2

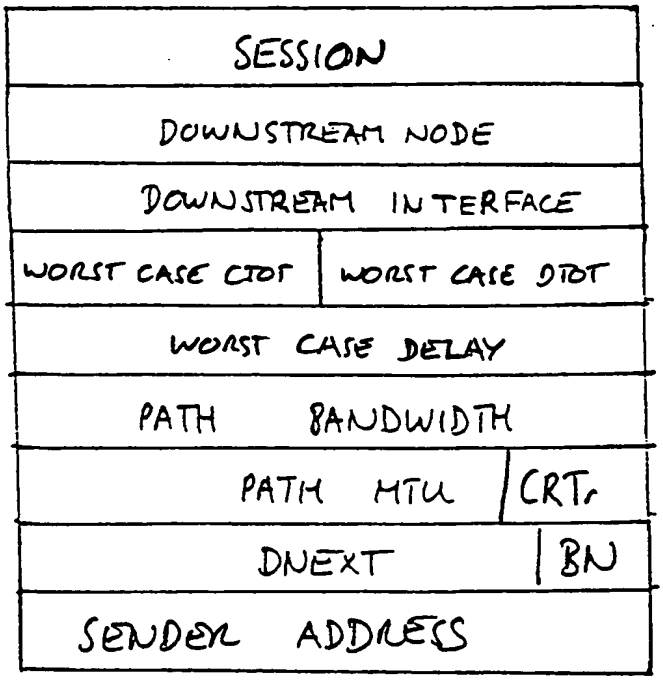


Figure 4

INTERNATIONAL SEARCH REPORT

International Application No
PCT/GB 98/03274

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 H04L29/06 H04L12/56

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 6 H04L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	SHENKER S ET AL: "TWO ISSUES IN RESERVATION ESTABLISHMENT" COMPUTER COMMUNICATIONS REVIEW, vol. 25, no. 4, 1 October 1995, pages 14-26, XP000541647	1-6
A	see paragraph 2.1 - paragraph 2.3 see paragraph 3.1.2 see paragraph 3.3 --- -/--	7

[X] Further documents are listed in the continuation of box C.

[ ] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
12 March 1999	26/03/1999

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Dupuis, H

Form PCT/ISA/210 (second sheet) (July 1992)